Brown, Katherine and Moreton, Joanna and Malla, Sunir and Aboobaker, A. Aziz and Emes, Richard D. and Tarlinton, Rachael E. (2012) Characterisation of retroviruses in the horse genome and their transcriptional activity via transcriptome sequencing. Virology, 433 (1). pp. 55-63. ISSN 0042-6822

Characterisation of Retroviruses in the Horse Genome and their Transcriptional Activity via Deep Sequencing.

Katherine Brown, Richard Emes, Rachael Tarlinton

Abstract

The recently released draft horse genome is incompletely characterised in terms of its repetitive element profile. This paper presents a characterisation of the endogenous retrovirus (ERVs) of the horse genome based on a data-mining strategy using murine leukaemia virus proteins as queries. Nine hundred and seventy eight ERV gene sequences were identified. Sequences were identified from the gamma, epsilon and beta retrovirus genera. At least one full length gammaretroviral locus was identified, though the gammaretroviral sequences are very degenerate. Using these data the RNA expression of these ERVs were derived from a deep sequencing data set from a variety of equine tissues. Unlike the well studied human and murine ERVs there do not appear to be particular phylogenetic groups of equine ERVs that are more transcriptionally active. Using this novel approach provided a more technically feasible method to characterise ERV expression than previous studies.

Introduction

Retroviruses that have been integrated into the DNA of a host's germ-line cells and have become heritable are known as endogenous retroviruses (ERVs). The life cycle of an ERV is initially identical to that of an exogenous retrovirus, involving reverse transcription of viral RNA into double stranded DNA, which is then integrated into the genome of the host and transcribed by cellular factors. ERVs are however gradually degraded over time by mutation, eventually losing their ability to replicate. Once a retrovirus has become endogenous, it behaves like any other genetic element, so it is subject to selection, mutation and genetic drift and can spread through the host population to fixation, or be eliminated from the population entirely (Jern and Coffin, 2008).

Until recently, the majority of endogenous retroviruses were identified using laboratory-based techniques (Gifford and Tristem, 2003). However, as more whole genome sequences become available, bioinformatics techniques are being increasingly used to characterise ERVs. Successful large-scale screening projects have been carried out on the human (Tristem, 2000; Villesen et al., 2004), chimpanzee (Polavarapu et al., 2006), mouse (Baillie et al., 2004; McCarthy and McDonald, 2004), rat (Baillie et al., 2004) and cow (Garcia-Etxebarria and Jugo, 2010) genomes, amongst others. In addition all genomes in the University of Southern California (UCSC) genome database (Kent et al., 2002) have a RepeatMasker track with basic annotation and classification of repetitive sites, including retroviral long terminal repeats (LTRs) (Smit et al., 1996).

Horses have long been important to human economic activity as a means of transport and haulage and still serve this function in many parts of the world. Even in developed countries horses are of considerable economic importance due to the popularity of riding as a sport and the gambling associated with this. A reference genome for the horse has been available since 2009 (Wade et al.,

2009). Forty seven percent of this genome is comprised of repetitive elements, including ERVs. Adelson et al. (2010) identified 389 ERV-like LTR containing elements based on RepeatMasker annotation, making up about 6% of the horse genome (Adelson et al., 2010). A large number of these elements are probably very short degenerate elements or solo LTRs without retroviral gene coding sequences, as is the case in other species, such as the cow (Garcia-Etxebarria and Jugo, 2010). Van der Kuyl (2011) performed a more detailed analysis of ERVs in the horse genome, using TBLASTN (Altschul et al., 1990) to compare the translated draft horse genome to retroviral *pol* gene fragments. Two hundred gamma- and beta- retrovirus like sequences were identified (van der Kuyl, 2011). One was characterised in detail, equus endogenous retrovirus EqERV-beta1, an intact beta-like retrovirus (van der Kuyl, 2011). The gammaretroviral sequences were not characterised further in this study and the BLAST based method utilised is known to give a very conservative estimate of the number of retroviral insertions (Garcia-Etxebarria and Jugo, 2010), so it is likely that there are further groups of retroviruses within the equine genome.

Widespread ERV expression at an RNA level is known to occur in most genomes studied to date (Denner, 2010). In some cases retroviral genes are known to have been co-opted into producing proteins essential in their hosts function, such as the HERV-W protein that encodes the human placental fusion gene syncytin (Mi et al., 2000) but this is a rarity. Whether expression of HERV RNA is merely the result of the escape from transcriptional suppression of "junk DNA" or serves a wider function in terms of RNA level control of transcription and translation is an unresolved issue. Certain groups of HERV loci, including HERV-K(HML-2), HERV H and HERV-W are frequently reported as being overexpressed in stem cell, germ cell, placental and neuronal tissues in humans (Antony et al., 2011; Flockerzi et al., 2008; Stauffer et al., 2004). Altered HERV expression has also been linked to schizophrenia, multiple sclerosis and a variety of cancers (Jern and Coffin, 2008) with varying degrees of robustness.

In many studies of HERV expression, although transcripts of a particular ERV lineage are quantified, there is no attempt to establish which ERV locus is being transcribed. Reference sequences are often selected to represent a particular lineage, however these are not necessarily the most active elements in that lineage (Oja et al., 2007). A few studies have attempted to assign expressed ERV transcripts to loci using varying methods. Oja et al. (2007) and Stauffer et al. (2004) both utilised the expressed sequence tag (EST) datasets available through public repositories. Oja et als relatively comprehensive study demonstrated that the majority of HERVs are inactive with 60% of HERV activity attributed to the 10 most active loci. As most of the HERVs did not include retroviral ORFs, they are clearly not performing viral functions and it was suggested that retroviral sequences have been co-opted by the host for other purposes (Oja et al., 2007). The proportion of active HERVs varied greatly between HERV lineages with for instance 20% of HERV-K elements being active as opposed to 2% of HERV-H elements (Oja et al., 2007).

RT-PCR methods have also been utilised. Flockerzi et al. (2008) looked at the expression of specific HERV-K (HML-2) loci in normal and malignant brain, breast and testicular tissue using sequencing of RT-PCR products of HML-2 specific primers to map specific transcripts back to the human genome. They located 23 transcriptionally active loci with one shown to be overexpressed in

testicular cancer. This method was more successful than using the expressed sequence tag database, however low level transcription was still sometimes missed, and polymorphic loci made it more difficult to assign transcripts to a locus (Flockerzi et al., 2008). Some studies have also utilised HERV specific microarrays to examine the issue of HERV RNA expression. These are however limited by the number of elements printed on the microarray (Diem et al., 2012; Gimenez et al., 2010; Haupt et al., 2011).

Next generation sequencing of whole transcriptomes and mapping of specific loci onto reference genomes presents an ideal way to overcome some of the limitations of the methods currently used to characterise ERV expression. This method allows comprehensive coverage of all the ERVs present within a sample without labour intensive cloning efforts or pre-selection of which loci will be examined. The accuracy of the platform utilised for mapping reads back to the genome may affect the certainty of localisation of reads to specific loci but this still represents a significant improvement on previous methods which were largely reliant on non-specific Q-PCR methods of detecting groups of ERVs.

This study presents data on a complete characterisation of the ERVs present in the equine genome using a data-mining strategy based on the Exonerate algorithm (Slater and Birney, 2005) with retroviral gene query sequences. This data set has been combined with a deep sequencing data set from RNA derived from horse tissues to examine the expression of individual loci at a tissue specific level. This methodology proved effective in identifying transcriptionally active loci in horse tissues.

Results

Genome Screening

A total of 978 DNA sequences representing potential ERVs could be identified in the horse genome (Supplementary Table 1). Of these, 112 were *gag* sequences, 813 were *pol* and 53 were *env*. Sequences flanked by pairs of LTR-like sequences, located using LTR harvest (Ellinghaus et al., 2008), were then identified. Sequences are denoted here as Ec2_chromosome_start position_end position, as suggested in Jern et al. (2005).

An 8427 base pair region, Ec2_chr5_27325585_27334012, was identified on the antisense strand of chromosome five, representing a relatively intact retroviral locus with the expected retroviral structure of LTR-*gag*-*pol*-*env*-LTR was identified (Figure 1). An additional region on chromosome 17, Ec2_chr17_37526914_37532405, was identified with all three viral coding regions but lacking LTRs. The locus on chromosome five was characterised in detail and open reading frames were detected in the protein coding region, although all were incomplete (Figure 1).

The two LTRs flanking this locus are identical except for three mutations, assuming the horse neutral substitution rate is somewhere between that of the mouse (2.2×10^{-9} substitutions per site per year) (Chinwalla et al., 2002) and the human (4.5×10^{-9} substitutions per site per year) (Lander et al., 2001), this gives an estimated integration date of 740741 to 1515152 years ago.

Sixty nine regions contained pairs of ERV genes in the expected orientation, 22 of which were flanked by paired LTRs (Table 1). The remaining 779 regions contained a single ERV-like element, 71 of which were flanked by paired LTRs (Table 1). Loci with LTRs are marked individually in Suplementary Table 1

BLAST searches and phylogenetic analysis identified gammaretrovirus like *gag* (112 sequences), *pol* (379 sequences) and *env* (53 sequences) genes (Supplementary Table 1). Only *pol* genes were identified in the betaretrovirus (41 sequences) and epsilonretrovirus (389 sequences) genera (Supplementary Table 1). No loci were identified in the lentivirus, spumavirus, alpharetrovirus or deltaretrovirus genera.

Phylogenetic trees (Figures 2 - 6) were generated to show the relationships between horse sequences and known members of each retroviral genus. The horse sequences were combined into consensus sequences and for the larger *pol* datasets, single sequences were chosen to represent well-supported clades.

Figures 2, 3 and 4 show that no gammaretroviral genes were identified falling within the clades with known active exogenous gammaretroviruses such as murine leukaemia virus, feline leukaemia virus and reticuloendotheliosis virus. Instead, the sequences clustered with the results of genome screening projects in the human, chimpanzee (Polavarapu et al., 2006) and cow (Garcia-Etxebarria and Jugo, 2010).

As shown in Figure 5, our method successfully identified 10 *pol* genes (represented by the sequence POL_beta_3) corresponding to EqERV described in van der Kuyl (2011). One of these was in the same position as the intact locus identified in van der Kuyls paper. Our phylogenetic analysis, placing this in a group with mouse mammary tumour virus (MMTV), also corresponds with van der Kuyls analysis. Three other groups of betaretrovirus were identified, one of which is similar to the sheep endogenous retrovirus 2 identified in the sheep genome by Klymuik et al. (2003). The remaining two betaretrovirus consensus sequences are distinct from other known betaretroviruses.

Unexpectedly, many ERVs identified in the horse cluster with the epsilonretroviruses (Figure 6). A large group of insertions was identified which clusters consistently with the human epsilon-like retrovirus identified by Jern et al. (2005). A second group clusters robustly with the *Dendrobates vertrimaculatus* ERVs (Tristem et al., 1996) although the two groups are distinct. The distinction between the gamma- and epsilon-like retroviruses was difficult to establish in several cases. One group of gammaretroviruses previously identified in the chimpanzee (Polavarapu et al., 2006) appears in our analysis to be epsilon- rather than gamma-like.

Analysis of Expression Data

Seventy nine of the 842 sequences (9.3%) identified on known chromosomes (those in unclassified DNA were excluded) had an expression level above one read per kilobase of exon per million mapped reads (RPKM). The representative sequences for these loci are underlined in Figures 2 to 6 and their maximum RPMK in any tissue provided in Supplementary Table 1. Only gamma and epsilon retroviruses had expression levels above this threshold. Expressed sequences were scattered throughout the phylogenetic tree with no particular groups of ERVs uniformly or highly expressed

compared with others. In three cases, *gag* and *pol* sequences from the same locus were expressed together, however in the majority of cases expression of genes from the same loci did not occur. There was no expression for the complete locus on chromosome 5 or of the known intact betaretrovirus locus.

Of those loci that were expressed at RPMK>1 a higher proportion (25.8%) were located either internal to or within 10 kB of an identified gene than the proportion in the entire dataset (8.5%)

Discussion

This study has expanded and characterised the known retroviral loci within the horse genome. The previously identified betaretroviral loci (van der Kuyl, 2011) were also identified here, though our search strategy using gammaretroviral search queries did not locate the *gag* or *env* genes of the full length locus identified in van der Kuyls paper. In both van der Kuyl's study and ours, the remaining gamma and beta retroviral sequences identified were fragmentary and difficult to classify. We have however categorised four new groups of betaretroviruses, 41 new groups of epsilonretroviruses and at least 53 new groups of gammaretroviruses in the horse genome. Our search strategy has clearly identified more loci than the BLAST based strategy used by van der Kuyl, however BLAST based strategies have been shown previously to be very conservative in identifying ERV loci (Garcia-Etxebarria and Jugo, 2010). The reduced number of *gag* and *env* compared to *pol* genes is also typical of ERVs in mammalian genomes and is attributed to the propensity for ERVs to retain functional *pol* genes, required for intercellular transposition longer than they retain the capsid and envelope genes necessary for intracellular infection (Bannert and Kurth, 2006).

Interestingly, a large number of the loci identified in this study clustered with the epsilon genus on phylogenetic analysis. Most of the epsilon retroviruses identified to date have been from fish, amphibian and reptile species (Kambol et al., 2003; Sinzelle et al., 2011; Tristem et al., 1996), with the exception of a human epsilon-like retrovirus identified by Jern et al. (2005). The identification of a number of epsilon-like sequences in the horse genome would indicate that these are more widespread in mammalia than previously thought. A number of clades did not cluster consistently with the gamma or epsilon genera but were intermediate between the known exogenous viruses in these two clades. This suggests that the gamma and epsilon genera may not be completely distinct phylogenetic groups.

As the horse ERV sequences identified here are from known positions on the horse genome, it was relatively straightforward to retrieve their RNA transcription levels from a deep sequencing data set. While confined to a small number of animals and tissues, statistically robust identification of transcription of individual loci could be readily achieved. This method allows greater accuracy in describing transcription of individual ERV loci with a far less labour intensive methodology than previously described methods. Indeed in our hands the most commonly used methodology for analysing ERV expression, quantitative PCR (QPCR), could not identify individual ERV loci with any certainty (data not shown). Using deep sequencing for transcriptome analysis also allowed global analysis of all ERV groups simultaneously, avoiding the selection bias towards more phylogenetically

interesting groups of ERVs evidenced in many studies on HERV expression and the limitations of the number of sequences that can be placed on a microarray chip.

While a number of ERVs are expressed at RPMK >1 in the tissues examined here, there are no distinctive groups of active ERVs with expressed ERVs scattered across the phylogenetic tree. A limitation of this study is that neurological and reproductive tissue which in other species do display distinctive ERV expression profiles (Flockerzi et al., 2008; Palmarini et al., 2004) were not available for analysis. Similarly to Oja et als EST study of HERVs, a large proportion of our equine ERVs appear to be transcriptionally silent. The increased proportion of active ERVs that are located either within or in close proximity to annotated genes when compared with the entire data set would indicate that a certain percentage of equine ERV transcriptional activity is related to expression of linked gene sequences. As very few of the ERV loci identified in this study are capable of producing anything resembling a functional protein clearly their expression is not linked to the production of ERV proteins. In this context it should be noted that the RNA data set utilised here was not mRNA enriched, this was deliberate to capture expression of any small or non-coding RNA not destined for protein production. Determining if particular ERVs are associated with expression in certain tissues or disease syndromes will require follow up studies with larger numbers of samples but this methodology has a much greater chance of providing definitive answers as to whether there is distinctive expression of ERV loci associated with certain disorders than previous approaches.

Materials and Methods

Genome Screening and Data Analysis

The Exonerate algorithm (Slater and Birney, 2005) was run sequentially on each chromosome of the horse genome (EquCab2) using full-length *gag*, *pol* and *env* genes from Moloney murine leukaemia virus (Shinnick et al., 1981) as query sequences. The cut-off for potential ERV genes was a length of 200 amino acids without exons. When predicted genes overlapped, the gene with the highest Exonerate score was selected (pipeline scripts to conduct this analysis are freely available from the authors at https://sites.google.com/site/emesbioinformatics/group-software). Potential amino acid sequences were generated for these sequences by selecting the translated query sequence producing the most significant hit in a BLASTX (Altschul et al., 1990) search against a database of amino acid sequences of known retroviruses.

Chromosome regions potentially encoding more than one endogenous retroviral gene were identified using an automated search of the output. ERV regions were taken as regions in which the start site of more than one retroviral gene was identified within an 8000 base pair region. LTR harvest (Benachenhou et al., 2009) was used to locate paired LTR-like regions and ERV regions with an LTR region starting within 5000 base pairs of each end were identified.

Amino acid sequences were aligned using MAFFT v6.824 (Katoh et al., 2002) under the FST-NS-2 algorithm and assigned to a genus using a phylogenetic tree generated using FastTree 2 (Price et al., 2010). Sequences assigned to each genus were aligned using the g-INS-i model in MAAFT and an identity matrix was generated for each. ERV sequences overlapping by more than 30 (*pol*) or

15 (*gag* and *env*) amino acids and sharing 75% sequence identity in the overlapping region (excluding gaps) were grouped and aligned using MUSCLE (Edgar, 2004), under the default settings, then manually adjusted. Consensus sequences were generated for these groups using the cons function of EMBOSS (Rice et al., 2000). For the large *pol* dataset, consensus sequences and the remaining sequences not incorporated into a consensus were realigned using the g-INS-i model and used to generate phylogenetic trees in FastTree, and where sequences not included in a consensus formed a monophyletic clade with branch support of greater than 75%, as calculated through FastTree using the Shimodaira and Hasegawa model (Shimodaira and Hasegawa, 1999), a single representative sequence was selected. The finalised datasets for each gene and genus were aligned with representative retroviruses (listed in Supplementary Table 2) under the g-INS-i model and trees were generated using PhyML, under the JTT amino acid substitution model, with optimised tree topology and between site variation. Branch support was calculated through PhyML using the aLRT model. The appropriate gene of equine infectious anaemia virus, a lentivirus, was used as an outgroup.

Deep Sequencing

Samples

Five tissue samples (kidney, jejunum, liver, spleen and mesenteric lymph node) were collected from an aged gelding (castrated male horse) euthanased due to osteoarthritis. These tissue blocks were stored in RNA later at -20 degrees until extraction. The tissue samples listed were collected from an animal euthanised for clinical reasons, by the veterinary surgeon, underthe Veterinary Surgeons act of 1966. Full informed consent of the owner was obtained for use of the samples, taken from that animal post-mortem. Lymphocytes isolated by Ficoll Paque (GE healthcare) from a healthy 11 year old welsh mountain pony gelding were kindly provided by Dr Julia Kydd (School of Veterinary Medicine and Science, University of Nottingham) under a the Home Office and local Ethical Approval Committee (PPL 40/3354). RNA extraction on these samples was performed using the Nucleospin RNA II mini kit (Machery Nagel) according to manufacturer's instructions. RNA from lymphocytes isolated from a healthy thoroughbred mare (the same horse whose DNA the horse genome is derived from) was kindly provided by Donald Miller (Baker Institute of Animal Health, Cornell University, USA) This horse was maintained atthe Baker Institute for Animal Health, Cornell University, Ithaca, N.Y.,USA. Animal care and research activities were performed in accordancewith the guidelines set forth by the Institutional Animal Care and UseCommittee of Cornell University, protocol # 1986-0216, approved throughMarch 2013.

Preparation and sequencing of SOLiD 3 whole transcriptome libraries

Approximately 10µg of total RNA extracted from each horse tissue sample was treated with 2U of Turbo DNase in 1 x DNase buffer (Ambion, cat. No. AM2238). Samples were further extracted using phenol pH 4.3: chloroform (50:50; Sigma, Cat. No. P4682-100 ML) and alcohol precipitated overnight at - 80℃. The RNA pellet was washed with 70% ethanol, air dried and resuspended in 10ul of

nuclease free water. Ribosomal RNA was removed with the Ribominus Eukaryotic kit (Invitrogen, Cat. No. A10837-08) as stated by the manufacturer. SOLiD whole transcriptome libraries were made as outlined in the Solid Whole transcriptome kit protocol (Applied Biosystems, Cat. No. 4425680). Library concentrations were determined with the Quant-it HS dsDNA assay kit (Invitrogen, Cat. No. Q32851). Sequencing was performed on a SOLiD 3 ABi sequencer to generate 50bp reads according to the manufacturer's instructions.

Read Mapping and calculation of relative expression values.

Raw reads were mapped in colour space to the reference genome (EquCab2). Annotations and position of known genes were obtained from Ensembl. Predicted ERVs were added to the annotation by generation of a modified GTF file. Reads were mapped using the CLC Genomics workbench (CLC Bio) mapping tool with a maximum of 2 mismatches allowed. Only uniquely mapped reads were processed for calculation of relative expression scores (RPKM) RPKM expression data was incorporated into the expanded phylogenetic trees described above.

Acknowledgements

Dr Julia Kydd (School of Veterinary Medicine and Science, University of Nottingham) and Dr Donald Miller (Baker Institute of Animal Health, Cornell University, USA) for their kind donation of lymphocyte samples and RNA. Dr's Aziz Aboobakar and Jo Moreton (Deep Seq, Centre for Genetics and Genomics, University of Nottingham) for the preparation and processing of the samples for the deep sequencing data set. Funding for this project was provided by the University of Nottingham. The funding body had no role in the execution and analysis of this study.

References

Adelson, D.L., Raison, J.M., Garber, M., Edgar, R.C., 2010. Interspersed repeats in the horse (Equus caballus); spatial correlations highlight conserved chromosomal domains. Animal Genetics 41, 91-99.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403-410.

Antony, J.M., DesLauriers, A.M., Bhat, R.K., Ellestad, K.K., Power, C., 2011. Human endogenous retroviruses and multiple sclerosis: innocent bystanders or disease determinants? Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1812, 162-176.

Baillie, G.J., van de Lagemaat, L.N., Baust, C., Mager, D.L., 2004. Multiple Groups of Endogenous Betaretroviruses in Mice, Rats, and Other Mammals. J. Virol. 78, 5784-5798.

Bannert, N., Kurth, R., 2006. The Evolutionary Dynamics of Human Endogenous Retroviral Families. Annual Review of Genomics and Human Genetics 7, 149-173.

Benachenhou, F., Jern, P., Oja, M., Sperber, G., Blikstad, V., Somervuo, P., Kaski, S., Blomberg, J., 2009. Evolutionary Conservation of Orthoretroviral Long Terminal Repeats (LTRs) and <italic>ab initio</italic> Detection of Single LTRs in Genomic Data. PLoS ONE 4, e5179.

Chinwalla, A.T., Cook, L.L., Delehaunty, K.D., Fewell, G.A., Fulton, L.A., Fulton, R.S., Graves, T.A., Hillier, L.D.W., Mardis, E.R., McPherson, J.D., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-562.

Denner, J., 2010. Endogenous Retroviruses, in: Kurth, R., Bannert, N. (Eds.), Retroviruses: Molecular Biology, Genomics and Pathogenesis. Caister Academic Press, Norfolk, UK, pp. 35-70.
Diem, O., Schäffner, M., Seifarth, W., Leib-Mösch, C., 2012. Influence of Antipsychotic Drugs on Human Endogenous Retrovirus (HERV) Transcription in Brain Cells. PLoS ONE 7, e30054.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792-1797.
Ellinghaus, D., Kurtz, S., Willhoeft, U., 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9, 18.
Flockerzi, A., Ruggieri, A., Frank, O., Sauter, M., Maldener, E., Kopper, B., Wullich, B., Seifarth, W., Muller-Lantzsch, N., Leib-Mosch, C., Meese, E., Mayer, J., 2008. Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project. BMC Genomics 9, 354.
Garcia-Etxebarria, K., Jugo, B.M., 2010. Genome-Wide Detection and Characterization of Endogenous Retroviruses in Bos taurus. Journal of Virology 84, 10852-10862.
Gifford, R., Tristem, M., 2003. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26, 291-315.
Gimenez, J., Montgiraud, C., Pichon, J.P., Bonnaud, B., Arsac, M., Ruel, K., Bouton, O., Mallet, F., 2010. Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control. Nucleic Acids Research 38, 2229-2246.
Haupt, S., Tisdale, M., Vincendeau, M., Clements, M.A., Gauthier, D.T., Lance, R., Semmes, O.J., Turqueti-Neves, A., Noessner, E., Leib-Mösch, C., 2011. Human endogenous retrovirus transcription profiles of the kidney and kidney-derived cell lines. Journal of General Virology 92, 2356-2366.
Jern, P., Coffin, J.M., 2008. Effects of Retroviruses on Host Genome Function. Annual Review of Genetics 42, 709-732.
Jern, P., Sperber, G., Blomberg, J., 2005. Use of Endogenous Retroviral Sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology 2, 50.
Kambol, R., Kabat, P., Tristem, M., 2003. Complete nucleotide sequence of an endogenous retrovirus from the amphibian, Xenopus laevis. Virology 311, 1-6.
Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30, 3059-3066.
Kent, W.J., 2002. BLAT—The BLAST-Like Alignment Tool. Genome Research 12, 656-664.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D., 2002. The human genome browser at UCSC. Genome Research 12, 996-1006.
Klymiuk, N., Muller, M., Brem, G., Aigner, B., 2003. Characterization of endogenous retroviruses in sheep. Journal of Virology 77, 11268.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.
McCarthy, E., McDonald, J., 2004. Long terminal repeat retrotransposons of Mus musculus. Genome Biology 5, R14.
Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.Y., Edouard, P., Howes, S., 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403, 785-789.
Oja, M., Peltonen, J., Blomberg, J., Kaski, S., 2007. Methods for estimating human endogenous retrovirus activities from EST databases. BMC Bioinformatics 8, S11.
Palmarini, M., Mura, M., Spencer, T.E., 2004. Endogenous betaretroviruses of sheep: teaching new lessons in retroviral interference and adaptation. Journal of General Virology 85, 1-13.

Polavarapu, N., Bowen, N.J., McDonald, J.F., 2006. Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. Genome Biol 7, R51.
Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5, e9490.
Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16, 276-277.
Shimodaira, H., Hasegawa, M., 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Molecular Biology and Evolution 16, 1114.
Shinnick, T.M., Lerner, R.A., Sutcliffe, J.G., 1981. Nucleotide sequence of Moloney murine leukaemia virus. Nature 293, 543-548.
Sinzelle, L., Carradec, Q., Paillard, E., Bronchain, O.J., Pollet, N., 2011. Characterization of a Xenopus tropicalis Endogenous Retrovirus with Developmental and Stress-Dependent Expression. J. Virol. 85, 2167-2179.
Slater, G., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31.
Smit, A., Hubley, R., Green, P., 1996. RepeatMasker Open-3.0.
Stauffer, Y., Theiler, G., Sperisen, P., Lebedev, Y., Jongeneel, C.V., 2004. Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues. Cancer Immun 4, 2.
Tristem, M., 2000. Identification and Characterization of Novel Human Endogenous Retrovirus Families by Phylogenetic Screening of the Human Genome Mapping Project Database. J. Virol. 74, 3715-3730.
Tristem, M., Herniou, E., Summers, K., Cook, J., 1996. Three retroviral sequences in amphibians are distinct from those in mammals and birds. J. Virol. 70, 4864-4870.
van der Kuyl, A.C., 2011. Characterization of a Full-Length Endogenous Beta-Retrovirus, EqERV-Beta1, in the Genome of the Horse (Equus caballus). Viruses 3, 620-628.
Villesen, P., Aagaard, L., Wiuf, C., Pedersen, F., 2004. Identification of endogenous retroviral reading frames in the human genome. Retrovirology 1, 32.
Wade, C., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T., Adelson, D., Bailey, E., Bellone, R., 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326, 865-867.

Table 1

Structure	Total	LTRs
Gag Pol Env	2	1
Gag Pol	65	21
Pol Env	3	0
Gag Env	1	1
Gag	42	3
Pol	692	65
Env	45	3

Table showing the number of multi gene containing loci identified along with whether the number of these loci flanked by LTR's